Ann. Occup. Hyg., Vol. 53, No. 2, pp. 173–180, 2009

doi:10.1093/annhyg/men085

Multimodel Inference and Multimodel Averaging in Empirical Modeling of Occupational Exposure Levels

J. LAVOUÉ* and P. O. DROZ

Department of Work Environment, Institute for Work and Health, Universities of Lausanne and Geneva, Bugnon 19, Lausanne 1005, Switzerland

Received 15 July 2008; in final form 10 October 2008; published online 27 January 2009

Empirical modeling of exposure levels has been popular for identifying exposure determinants in occupational hygiene. Traditional data-driven methods used to choose a model on which to base inferences have typically not accounted for the uncertainty linked to the process of selecting the final model. Several new approaches propose making statistical inferences from a set of plausible models rather than from a single model regarded as 'best'. This paper introduces the multimodel averaging approach described in the monograph by Burnham and Anderson. In their approach, a set of plausible models are defined *a priori* by taking into account the sample size and previous knowledge of variables influent on exposure levels. The Akaike information criterion is then calculated to evaluate the relative support of the data for each model, expressed as Akaike weight, to be interpreted as the probability of the model being the best approximating model given the model set. The model weights can then be used to rank models, quantify the evidence favoring one over another, perform multimodel prediction, estimate the relative influence of the potential predictors and estimate multimodel-averaged effects of determinants. The whole approach is illustrated with the analysis of a data set of 1500 volatile organic compound exposure levels collected by the Institute for work and health (Lausanne, Switzerland) over 20 years, each concentration having been divided by the relevant Swiss occupational exposure limit and log-transformed before analysis. Multimodel inference represents a promising procedure for modeling exposure levels that incorporates the notion that several models can be supported by the data and permits to evaluate to a certain extent model selection uncertainty, which is seldom mentioned in current practice.

Keywords: exposure assessment; exposure determinant; linear models; model selection; multimodel inference

INTRODUCTION

The last two decades have seen extensive use of empirical statistical models to report summaries of exposure data sets in the industrial hygiene literature. These models typically attempt to establish statistical links between the measured exposure levels and individual and environmental variables documented at the time of measurement. They can be used to predict exposure for a particular combination of variables or identify influent predictors, the so-called exposure determinants. The most popular models in occupational hygiene are certainly linear multiple regression models (Burstyn and Teschke, 1999).

Most analyses of this kind involve a number of measurements complemented by a set of potential predictors, such as job, type of process or the presence/absence of ventilation. The output of the analysis is a subset of these variables being identified as determinants while the others are deemed non-influential, based on their presence or not in the final chosen model. The strategy leading to choose the final model has therefore a fundamental impact on the conclusions drawn from the analysis. A common feature of exposure determinant–exploration studies is that the modeler has usually no specific hypothesis to test but rather is looking for influent predictors in a group of plausible candidates. This is especially the case when analyzing exposure data banks (e.g. the National Exposure Database, NEDB, in the UK), for which the analyst has no control on the type of ancillary information accompanying measurement results. He therefore has to rely on some procedure measuring goodness-of-fit to the data that will help select a 'best' model. This is different for instance

*Author to whom correspondence should be addressed.
Tel: +1 (514) 890 8000 #15913; fax: +1 (514) 412 7106;
e-mail: jerome.lavoue@umontreal.ca

from some epidemiology studies in which the model is specified in advance by the researcher based on knowledge of disease mechanism and confounding agents.

Data-driven model selection, i.e. choice of a model based on some manipulation of the data, even for the simplest linear regression, is still an open research area. It is not the scope of this paper to review this topic, which is extensively covered, for example in Leinhart and Zucchini or McQuarrie and Tsai (Linhart and Zucchini, 1986; McQuarrie and Tsai, 1998). Briefly, the main issue is that data are used at the same time to help choose a model and estimate parameters. This tends to yield results that reflect the data at hand but are not robust to external validation, i.e. do not reflect the population of interest. Despite ongoing debate on particular techniques, there seem to be consensus that hypothesis testing-based (using *P*-values) stepwise methods, the most popular in applied fields, are also the most problematic and susceptible to lead to spurious relationships and underestimated uncertainty (Harrel, 2001). A typical example of such procedure would start from the model containing the variable with the lowest *P*-value in univariate analyses. Then all models containing that variable and one other would be fitted, and again the variable with the lowest *P*-value would be kept, and so on until no additional variable achieve the 0.05 cutoff. Variations of stepwise procedure include starting from a model with all variables and progressively removing the non-significant ones or testing addition and removal of variables at each step.

Another limitation of these approaches is that a single best model will be chosen in the end, regardless of how far the 'second best' model actually was in terms of performance. Indeed, it is possible that several competing models were quite close, perhaps close enough that the difference in the goodness-of-fit criterion did not represent any meaningful evidence. In addition to loosing interesting information, it is quite possible, in circumstances of several models close to one another, that another sample of data would have yielded another best model. This uncertainty, linked to the modeling strategy rather than parameter estimation, is seldom mentioned in published analyses.

In the last decade, a new approach attempting to address the issue of modeling uncertainty was described coming from the field of Bayesian statistics. The principle involves averaging predictions across a set of models defined a priori with a weighting associated to their quality, which is expressed as the probability of being the best model (Raftery *et al.*, 1997). The estimated probabilities sum to 1 across the model set and can be used to appraise the amount of evidence in favor of a model (or a subset of models) compared to another. In particular, the 'distance' between the best model and the next ones can be quantified. For example, in a set of 100 models, two models could have probabilities of 0.50 and 0.45, with the others with much lower probabilities summing to 0.05. The analyst would conclude that the two models represent the majority of the evidence (i.e. 95%), but that there is no reason to choose one over the other: the model with the highest probability being only $0.50/0.45 = 1.1$ times more likely to be the best model than the one with the second highest. The procedure, labeled Bayesian model averaging (BMA), has since found use in the field of environmental epidemiology, for the derivation of dose–response relationship (Clyde, 2000; Martin and Roberts, 2006), but not, in our knowledge, in the field of occupational exposure assessment. The development on BMA methods has been somewhat hampered by the computer intensive calculations required, the lack of implementation in standard statistical packages and of widespread availability of the required training.

More recently, Burnham and Anderson have proposed a procedure similar to but simpler than BMA for calculating model weights (Burnham and Anderson, 2002, 2004). In their method, so-called 'Akaike weights' are calculated for each model, being interpreted as the probability of being the best approximating model given the data at hand and the model set. Burnham and Anderons's approach to multimodel inference, initially proposed in the field of ecology (Poeter and Anderson, 2005; Hollister *et al.*, 2008), also appeared in economy (Hansen, 2007), genetics (Posada and Buckley, 2004; Abdueva *et al.*, 2006), social sciences (Burnham and Anderson, 2004) and psychology (Wagenmakers and Farrell, 2004), and Moon *et al.* have proposed a variation in the field of risk analysis for estimating effective doses for microbial infection (Moon *et al.*, 2005). In addition, like BMA, to attempting to take into account modeling uncertainty, the approach presented by Burnham and Anderson can be readily implemented using standard statistical packages.

The main objective of this paper is to advocate the use of methods that explicitly account for the uncertainty linked to model selection in data-based empirical modeling in the field of occupational exposure assessment. To this end, we describe the simple and intuitive framework proposed by Burnham and Anderson and illustrate its use with a simplified analysis of retrospective multi-industry exposure measurements collected over the years at the Institute for Work and Health. We begin by presenting the illustrating data set and the linear model framework then proceed to describing the multimodel method alongside its practical implementation with the data set.

MODELING DATA SET

Since 1986, the Institute for Work and Health has maintained an exposure database containing all measurements made by its hygienists in various workplaces. There are currently ~8500 measurements in the data bank including biomonotoring results, volatile organic compounds (VOCs) and dust measurements. Information accompanying each measurement includes sampling and analytical method, date of sampling, type of measurement (source sampling, general area, personal...), sampling time, industry and job as coded by the Swiss Federal Institute of Statistics (OFS), origin of the measurement (request from private companies, in-house research projects) and a code identifying the person having measured.

For the purpose of this study, we set to explore the extent to which variables in the IST database could explain the variations of the recorded VOC concentrations. Since no single agent had enough data to permit such kind of analysis, we pooled all agents and standardized the concentrations by their Swiss 8-h occupational exposure limit (OEL) (each concentration was divided by the relevant exposure limit), thus creating a 'compliance index'. After cleaning of the data set, we kept for analysis only data corresponding to VOCs with an OEL in the list of OELs enforced in Switzerland by the Swiss Accident Insurance Fund (SUVA) ($n = 2588$). Further restriction to sample duration between 30 min and 12 h and to combinations of industry and job with at least 10 data yielded personal and area data sets of 698 and 722 measurements, respectively.

LINEAR MODELS FOR OCCUPATIONAL EXPOSURE

We modeled the log-transformed standardized concentrations as a linear function of other variables using the linear models framework (Neter *et al.*, 1996). The chosen model structure can be described using equation (1). This structure has been the most frequently used in the literature for reporting relationships between occupational exposure levels and determinants (Burstyn and Teschke, 1999).

$$\mathrm{Ln}(C_i) = \left(\sum_{f=1}^{p} (\text{Fixed effects})_f \right)_i + (\text{Error})_i, \quad (1)$$

$i = 1,\ldots,M$ where there are M measurements and p fixed effects included in the model. $\mathrm{Ln}(C_i)$ is the natural logarithm of the ith standardized concentration. The main model assumption is that (Error) follows a normal distribution independent of the fixed effects. As an illustration, for a model with industry as a categorical predictor and year of sampling as a continuous predictor, equation 1 would translate to:

$$\mathrm{Ln}(C_i) = \beta_{\text{industry}(j)} + \beta_{\text{year}} \times \text{year} + (\text{Error})_i, \quad (2)$$

Where there are $j = 1,\ldots,K$ industry category. $\beta_{\text{industry}(j)}$ is the coefficient for the jth industry category, and β_{year} is the slope of the linear temporal trend.

IMPLEMENTING MULTIMODEL INFERENCE USING THE FRAMEWORK PRESENTED BY BURNHAM AND ANDERSON (BURNHAM AND ANDERSON, 2002)

Definition of the model set

The first step in implementing multimodel inference consists in defining a set of plausible models. This step is fundamental since all results of the multimodel analyses are conditional on the model set, i.e. all conclusions are drawn 'given the model set'. According to Burnham and Anderson, knowledge of the subject matter should play a considerable role in limiting the size of the initial model set. In particular, they recommend that analyses where more models are fitted than the sample size be regarded as exploratory. This situation can be easily met in cases where a number of variables are available, with no specific knowledge of which should be excluded or kept or which interaction terms should be considered. For example, testing all possible combinations of 10 variables would correspond to $2^{10} = 1024$ possible models, not counting potential interactions.

For our analysis, we selected the model set by taking into account two main constraints: the need to have a sufficient number of data per estimated parameter (we arbitrarily set a target at 10 data per parameter) and to limit the number of models relative to the sample size. Moreover, we selected an 'all combinations' scheme, i.e. all possible models for a set of variables without interactions, in order to be able to estimate multimodel-averaged effects of predictors and to quantify the relative importance of all variables compared to each other (see below for the constraints linked to each procedure). Table 1 presents the seven variables that were selected for the analysis along with descriptive statistics, representing 128 different models to be fit to the data (compared to ~700 data in both area and personal data sets).

The full model (i.e. containing all variables in Table 1) explained 40.5 and 36.8% of the variations of the log-transformed personal and area compliance indices, respectively. These results are similar to other multi-industry modeling studies (see for example, Teschke *et al.*, 1999) and show that the model set includes relevant predictors. The multimodel procedure is then going to help identify which submodels provide a good approximation of the data without containing superfluous predictors (i.e. parsimonious models).

Table 1. Variables tested in the empirical statistical models

Variable	Type/data per category		Description
Industry/job	Nominal (21 and 23 categories for personal and area data, respectively)		Combination of industry and job codes
Year	Continuous (integer)		Year of sampling
	1989 to 2006 (shifted to vary from 0 to 17)		
Season	Nominal (four categories)		Season of sampling as defined by the following cut-off dates: winter (12/22 to 3/20), spring (3/21 to 6/21), summer (6/22 to 9/22), autumn (9/23 to 12/21)
	1. Winter (261P, 182A)[a]		
	2. Spring (186P, 163A)		
	3. Summer (79P, 76A)		
	4. Autumn (164P, 207A)		
Sample type	Personal data set	Area data set	Localization of sampling for area measurements and sex of worker sampled for personal measurements
	1. Female (162)	1. Source sampling (108)	
	2. Male (494)	2. Close to source (307)	
	3. Not documented (34)	3. In room containing source (169)	
		4. Not in room containing source (11)	
		5. Ventilation exhaust (33)	
Reason	Nominal		Origin of measurement, either internal (in house research projects) or external (request from private companies)
	1. Internal (166P, 121A)		
	2. External (524P, 507A)		
Duration	Continuous (integer)		Sampling duration in minutes
	Interquartile interval (P) 83–300 min[b]		
	Interquartile interval (A) 73–300 min		
Volatility	Nominal		Volatility of compound according to its vapor pressure at 20°C:
	1. Low (80P, 176A)		
	2. Medium (201P, 119A)		<12.5 mmHg: low
	3. High (184P, 134A)		12.5–69 mmHg: medium
	4. Gas (225P, 179A)		69–760 mmHg: high
	Low/medium and medium/high thresholds were taken as 33th and 66th percentiles of vapor pressures in the global data set		>760 mmHg: gas

[a]Number of data in the personal (P) and area (A) data sets.
[b]25th and 75th percentiles of sampling times in the personal and area data sets.

Selecting a performance criterion: introducing the Akaike information criterion and Bayesian information criterion

Recognition of the issues associated with the use of hypothesis tests to build models has led to the development of alternative procedures. Among the wide array of published methods, so-called 'information criteria' have enjoyed much popularity. These quantities are calculated for each model and allow their comparison with each other, the model with the lowest value being usually favored. The most widespread information criteria are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), the latter also known as the Schwartz information criterion (Kuha, 2004). AIC (equation 3) was derived by Akaike as an asymptotic estimator quantifying information loss when a model is used to approximate the truth. Minimizing AIC therefore minimizes the information loss over a set of models.

$$\mathrm{AIC} = -2\log(\ell(M|\ \mathrm{data})) + 2K, \qquad (3)$$

where $\ell(M|\ \mathrm{data})$ is the likelihood of the model given the data and K is the number of parameters of the model. $\ell(M|\ \mathrm{data})$ is a standard output of most statistical packages for numerous model structures. When the number of parameters of the largest model is such that $n/K < 40$, n being the sample size, Burnham and Anderson recommend the use of a modified version of AIC:

$$\mathrm{AIC.c} = -2\log(\ell(M|\ \mathrm{data})) + 2K + \frac{2K(K+1)}{n-K-1}. \qquad (4)$$

BIC (equation 5) was derived in a Bayesian framework as an approximation of quantities measuring the odds of a model being the true model given the data.

$$\mathrm{BIC} = -2\log(\ell(M|\ \mathrm{data})) + \ln(n) \times K. \qquad (5)$$

Both criteria include a measure of goodness-of-fit to the data (the likelihood) and a penalty term for the

number of parameters. For BIC, the number of parameters is more penalizing than for AIC, therefore models selected with BIC tend to contain fewer variables than those selected with AIC.

Calculating model weights

The second step of this procedure involves fitting each model to the data and calculating model weights (equations 1 and 2 in the supplementary data, available at *Annals of Occupational Hygiene* online) which, noted 'w_i', can be interpreted as the probability that the model is the best approximating model given the data at hand and the initial model set. While they can be calculated using different performance criteria (see for example, Hjort and Claeskens, 2003 or Hansen, 2007), Burnham and Anderson advocate the use of AIC or AIC.c (Burnham and Anderson, 2004).

The calculated model weights provide a way of ranking models in the set, and the weight of evidence favoring a model over another can be estimated with the ratio of their respective weights. Specific subsets of models can be compared to other subsets in the same way (e.g. all models containing an interaction term versus those not containing it). Using guidelines for evidence ratios quoted by Lukacs *et al.*, a ratio $\sim$10 corresponds to limited-to-moderate support while $>$100 is required to report 'strong support' (Lukacs *et al.*, 2007). The model weights can also be used to define a confidence set of models representing the majority of evidence. For example, if the sum of the five highest w_i values is $>$0.95, the corresponding models form a 95% confidence set, and one could decide to base inferences only on these five models.

Table 2 presents model weights calculated using AIC.c for the five personal and area models with the highest weights. As can be seen from Table 2, the models with the lowest AIC.c in the personal and area data sets are 1.6 (0.35/0.22) and 1.7 (0.46/0.28) times more likely to be the best fitting model than the models with the second lowest values. In both cases, the five models with the highest w_i values represent $\sim$90% (88 and 93% for personal and area data, respectively) of the evidence (i.e. the sum of their w_i is $\sim$90%). The model weights in Table 2 illustrate the kind of information that is lost when using a 'single-model' approach. Hence, the best fitting models for area and personal data represent $<$50% of the evidence and are only $\sim$1.5 times more likely to be the best approximating model compared to the second best models. There is therefore no overwhelming reason to choose them as the only useful models.

Identifying influential predictors

In empirical modeling of occupational exposure data, assessing the relative influence of variables has a major importance since it permits identifying exposure determinants. In traditional analyses, potential predictors tend to be declared either influential or not based on their presence in the single model chosen by the modeling strategy. The framework described by Burnham and Anderson allows quantifying the importance of variables relative to each other for a particular structure of the model set: the variables to be compared should be tested with a single parameterization (e.g. duration as a continuous or category variable), should be all included in the same number of models, should not be involved in interactions and the models should have the same basis (e.g. all linear models). When these conditions are met, the weights of the models containing variable x_j can be summed, yielding a quantity noted $W_+(j)$ and called relative importance weight for variable J. $W_+(j)$ can then be compared with $W_+(k)$ for another variable. As an illustration, for two variables J and K with $W_+(j) = 0.50$ and $W_+(k) = 0.25$, one might say that variable J is twice as important as variable K. These values therefore provide a measure of the weight of evidence supporting the presence of an actual relationship between a variable and the response relative to other variables tested (i.e. given the model set). Table 3 presents the $W_+(j)$ for the seven variables tested in this analysis, again calculated using AIC.c. For the personal data in Table 3, industry/job, year, duration and volatility appear clearly as the most important predictors, while season, sample type and reason have lower $W_+(j)$ values. For the area data, season and volatility have lower $W_+(j)$ values than other variables.

Table 2. Akaike weights for the five best personal and area models

	Personal data					Area data				
Industry/job	X	X	X	X	X	X	X	X	X	X
Year	X	X	X	X	X	X	X	X	X	X
Season		X		X		X		X		X
Sample type					X	X	X	X	X	X
Reason			X	X		X	X	X	X	X
Duration	X	X	X	X	X	X	X	X	X	
Volatility	X	X	X	X	X			X	X	
Akaike weight (w_i)	0.35	0.22	0.17	0.09	0.05	0.46	0.28	0.09	0.07	0.03

Table 3. Relative importance of the seven variables tested in the models

Personal	$W_+(j)$[a]	Area	$W_+(j)$
Industry/job	>0.99	Industry/job	>0.99
Year	>0.99	Year	>0.99
Season	0.37	Season	0.62
Sample type	0.13	Sample type	>0.99
Reason	0.31	Reason	0.96
Duration	0.95	Duration	0.94
Volatility	>0.99	Volatility	0.18

[a]Relative importance of variable *J*

Multimodel-averaged inference

The model weights also form the basis on which to perform inference, i.e. predicting the response for a specific set of conditions (e.g. job *A* during year *X* with a sampling duration of *Y*...) not conditional on a single model but over the whole model set. The predictions are made from each model and then averaged, each prediction being weighted using its w_i value. For each prediction, a variance can be estimated that is not conditional on a particular model and comprises within- and between-model (i.e. modeling uncertainty) components (see equation 3 and 4 in the supplementary data, available at *Annals of Occupational Hygiene* online). Based on limited simulation, Burnham found that confidence intervals calculated with these variance estimates generally were wider but more realistic than those obtained with traditional methods, which tended to yield less than desirable results.

We do not present here predictions of the compliance index for specific industries in the Swiss database because discussion of these predictions and detailed analysis of the usefulness of the IST data bank to reflect occupational exposure in Switzerland is outside the scope of the present paper. Moreover, the predictions would have reflected temporal trends we believe are mainly due to a time-changing selection bias in the database. We nevertheless report here that the unconditional standard errors for the personal predictions, calculated to estimate 95% confidence intervals, were between 5 and 20% higher than the conditional standard errors obtained by using only the best fitting model (between 2 and 12% for the area data). They illustrate the added uncertainty caused by including model selection as a process subject to variability.

Multimodel-averaged estimators of effects

Multimodel-averaged estimates of effects (i.e. model coefficients) can also be calculated with the model weights (see equations 5–7 in the supplementary data, available at *Annals of Occupational Hygiene* online). The calculation is similar to making predictions but requires the models to be linear models (such as those presented here) and the coefficients to keep the same interpretation across the model set or at least across the subset over which the averaging is to be performed.

In the particular case of 'all combinations' model sets, Burnham and Anderson propose another way of calculating multimodel coefficient estimates that takes into account the relative importance weight of the variable of interest. Hence, the coefficients are averaged over the whole model set, being taken as 0 for models not including the variable. Using this approach, the multimodel-averaged coefficient will be 'shrunk' toward zero compared to the previous calculations (see equation 8–10 in the supplementary data, available at *Annals of Occupational Hygiene* online). The extent of the shrinkage will depend on the cumulated weight of the models without the variable. This method of estimation is appealing because only averaging over models in which a variable is present is likely to yield to an upward model selection-related bias.

Multimodel-averaged estimates of effects of all variables but industry/job (which included >20 categories), shrunk and unshrunk are presented in the supplementary data (available at *Annals of Occupational Hygiene* online). We limit our presentation here to the effect of duration and volatility for the area data set for illustration purpose. For the area data, the compliance index was estimated to decrease by 14 and 15% with the shrunk and unshrunk methods, respectively, for a 50% increase in the sampling time. Both estimates are very close because duration had a relative importance weight very close to 1 (0.94). On the other hand, unshrunk estimates for volatility, with the 'gas' category at 100% exposure, were 'high' 183%, 'medium' 252% and 'low' 121%. Even with wide confidence intervals, these values, close to those obtained with the full model (i.e. containing all variables), would suggest an odd relationship between exposure and volatility. Because this variable had a low relative importance weight, the shrunk estimates are reduced to 97, 103 and 93%, almost equivalent to no effect. Burnham and Anderson underline the need for further work in this area of their approach (i.e. adequacy of $W_+(j)$ as shrinkage factor, exact variance of the shrunk estimator), but the general use of shrinkage to account for model selection bias (i.e. over estimation of effects due to the use of the same data set for model building and effect estimation) is well established in the statistical literature (Harrel, 2001).

Comparison between AIC, AIC.c and BIC

We compared the relative importance weights of all variables determined using AIC, AIC.c and BIC and also calculated the same weights using

a bootstrap procedure, described in details in the supplementary data (see also Table 2 of the supplementary data, available at *Annals of Occupational Hygiene* online). The bootstrap weights were for the most part very close to their 'analytical' counterpart. The weights based on BIC favored simpler models than AIC or AIC.c (which yielded close results), reflected in a smaller relative importance weight for most variables. AIC models are known to yield generally more complex models than BIC (Burnham and Anderson, 2004; Kuha, 2004). With regards to the AIC versus BIC question, we support the view of Kuha, who evocated the possibility, when AIC and BIC disagree to use the corresponding models as 'bounds for a range of acceptable models' (Kuha, 2004).

All analyses were conducted with versions 6.1 and 7.0 of the statistical software S-plus Professional Edition for Windows (Insightful Corp., Seattle, WA).

DISCUSSION

We begin this section by quoting what Chatfield regarded as his main message in a paper read before the Royal Statistical Society in 1995 (Chatfield, 1995): 'When a model is formulated and fitted to the same data, inferences made from it will be biased and overoptimistic when they ignore the data analytic actions which preceded inference. Statisticians must stop pretending that model uncertainty does not exist and begin to find way of coping with it'.

In this paper, we presented a methodology that attempts to include modeling uncertainty in the analysis in a simple and intuitive way and that has already found its way into applied work in other scientific fields. While we find Burnham and Anderson's approach appealing, we in no way advocate exclusive use of their proposition as a universal panacea to model selection issues. An obvious limitation is indeed the fact that results from such analyses, albeit unconditional on a particular model, are conditional on the model set, and the modeling uncertainty is therefore only approximated. There is still, and probably will continue to be debate on how empirical modeling should be approached, which multimodel method is most adequate and in particular whether AIC, BIC, other ML-based criterion or some form of Bayesian or bootstrap-based model averaging should be preferred (Guthery *et al.*, 2005; Richards, 2005; Link and Barker, 2006; Ward, 2008). What seems important to us, as underlined by Ye *et al.*, is that the principle of multimodel averaging is more crucial in improving reliability of the results than the question of which particular multimodelling approach to use (Ye *et al.*, 2008).

In conclusion, we believe empirical modeling studies in the field of occupational exposure assessment should attempt to account for modeling uncertainty. In this regard, multimodel averaging using the approach of Burnham and Anderson provides in our view an easy to implement and intuitive methodology.

SUPPLEMENTARY MATERIAL

Supplementary data can be found at http://annhyg.oxfordjournals.org/

Acknowledgements—The authors would like to thank Drs Pascal Wild and Michel Grzebyk, of the Institut national de recherche et de sécurité (INRS, France) for their very helpful comments on earlier drafts of the present paper.

REFERENCES

Abdueva D, Skvortsov D, Tavare S. (2006) Non-linear analysis of GeneChip arrays. Nucleic Acids Res; 34: e105.

Burnham KP, Anderson DR. (2002) Model selection and multimodel inference. 2nd edn. New York, NY: Springer Science+Business Media Inc.

Burnham KP, Anderson DR. (2004) Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res; 33: 261–304.

Burstyn I, Teschke K. (1999) Studying the determinants of exposure: a review of methods. Am Ind Hyg Assoc J; 60: 57–72.

Chatfield C. (1995) Model uncertainty, data mining, and statistical inference. J R Stat Soc; 158: 419–66.

Clyde M. (2000) Model uncertainty and health effects studies for particulate matter. Environmetrics; 11: 745–63.

Guthery FS, Brennan LA, Peterson MJ *et al*. (2005) Information theory in wildlife science: critique and viewpoint. J Wildl Manage; 69: 457–65.

Hansen BE. (2007) Least squares model averaging. Econometrika; 75: 1175–89.

Harrel FEJ. (2001) Regression modeling strategies—with applications to linear models, logistic regression, and survival analysis. New York, NY: Springer.

Hjort NL, Claeskens G. (2003) Frequentist model average estimators. J Am Stat Assoc; 98: 879–99.

Hollister JW, August PV, Paul JF *et al*. (2008) Predicting estuarine sediment metal concentration and inferred ecological conditions: an information theoretic approach. J Environ Qual; 37: 234–44.

Kuha J. (2004) AIC and BIC: comparisons of assumptions and performance. Sociol Methods Res; 33: 188–229.

Linhart H, Zucchini W. (1986) Model selection. New York, NY: John Wiley & Sons.

Link WA, Barker RJ. (2006) Model weights and the foundations of multimodel inference. Ecology; 87: 2626–35.

Lukacs PM, Thompson WL, Kendall WL *et al*. (2007) Concerns regarding a call for pluralism of information theory and hypothesis testing. J Appl Ecol; 44: 456–60.

Martin MA, Roberts S. (2006) Bootstrap model averaging in time series studies of particulate matter air pollution and mortality. J Expo Sci Environ Epidemiol; 16: 242–50.

McQuarrie ADR, Tsai CL. (1998) Regression and time series model selection. Hackensack, NJ: World Scientific Publishing Company.

Moon H, Kim HJ, Chen JJ *et al*. (2005) Model averaging using the Kullback information criterion in estimating effective doses for microbial infection and illness. Risk Anal; 25: 1147–59.

Neter J, Kutner M, Nachtsheim CJ *et al*. (1996) Applied linear statistical models. New York, NY: WCB McGraw-Hill.

Poeter E, Anderson D. (2005) Multimodel ranking and inference in ground water modeling. Ground Water; 43: 597–605.

Posada D, Buckley TR. (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol; 53: 793–808.

Raftery AE, Madigan D, Hoeting A. (1997) Bayesian model averaging for linear regression models. J Am Stat Assoc; 92: 179–91.

Richards SA. (2005) Testing ecological theory using the information-theoretic: examples and cautionary results. Ecology; 86: 2805–14.

Teschke K, Marion SA, Vaughan TL *et al*. (1999) Exposure to Wood Dust in U.S. Industries and Occupation, 1979 to 1997. Am J Ind Med; 35: 581–9.

Wagenmakers EJ, Farrell S. (2004) AIC model selection using Akaike weights. Psychon Bull Rev; 11: 192–6.

Ward EJ. (2008) A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. Ecol Model; 211: 1–10.

Ye M, Meyer PD, Neuman S. (2008) On model selection criteria in multimodel analysis. Water Resour Res; 44: W03428.